Estimating the Distribution Function and Predicting Observables

by

Seymour Geisser

University of Minnesota

Technical Report No. 361

November 1979

Estimating the Distribution Function and Predicting Observables*

Seymour Geisser
University of Minnesota

1. Introduction

The predictive approach in Statistics involves nothing more profound than inferring how future or potential observables will be distributed given certain assumptions. This may involve estimating the sampling distribution function itself or making direct or indirect probabilistic statements about regions for future observables. In classical statistical theory there is an apparent distinction between predicting the next observation in a series of i.i.d. random variables and estimating the distribution function. For example if $X_1, \ldots, X_N, X_{N+1}$ are i.i.d. $N(\mu,\sigma^2)$ then it is well known that a tolerance (predictive) interval for the next observation is obtained from the "student" random variable with $N - 1$ degrees of freedom, namely,

$$(1.1) \qquad t_{N-1} = \frac{X_{N+1} - \bar{X}}{s\sqrt{1 + N^{-1}}}$$

where $\bar{X} = N^{-1} \sum_{i=1}^{N} X_i$ and $(N - 1)s^2 = \sum_{i=1}^{N} (X_i - \bar{X})^2$. This yields

$$(1.2) \qquad \Pr\left[\bar{X} - t_\alpha s\sqrt{\frac{N}{N+1}} \le X_{N+1} \le \bar{X} + t_\alpha s\sqrt{\frac{N}{N+1}}\right] = 1 - 2\alpha$$

where t_α is the αth percentile point in the right tail of t_{N-1}. Hence the $1 - 2\alpha$ predictive tolerance interval for X_{N+1} is defined for $\bar{X} = \bar{x}$. On the other hand straightforward estimation of the sampling distribution would lead to

- - - - - - - - - - -

*Work was supported in part by NIH-GM-25271.

$$(1.3) \qquad \hat{N}(\mu, \sigma^2) = N(\hat{\mu}, \hat{\sigma}^2) = N(\bar{x}, s^2) ,$$

or

$$(1.4) \qquad N(\bar{x}, s^2(1 + \tfrac{1}{N})) .$$

However it is clear that if we used (1.3) or (1.4) the frequency property of the random interval covering a random variable no longer obtains. On the other hand, the known or assumed normality of the sampling distribution is preserved. Moreover, from another perspective, it has been shown by Murray(1977) that the use of the t_{N-1} distribution as the estimator enjoys the property of being the best estimator of the $N(\mu, \sigma^2)$ which is a function of $\bar{x}$ and s^2 in the following frequency sense: Let the Kullback-Leibler (1951) information measure be averaged over the sample space so that for $n(x;\mu,\sigma^2)$ denoting the normal density,

$$(1.4) \qquad M(\mu,\sigma^2) = E_{X_1,\ldots,X_N}\left[\int n(x; \mu, \sigma^2)\log \frac{f(x; \bar{x}, s^2)}{n(x; \mu, \sigma^2)}dx \right]$$

$$= E_{\bar{X},s^2}[K(\mu, \sigma^2, \bar{X}, s^2)] .$$

Then $M(\mu,\sigma^2)$ is minimized w.r.t. all densities depending on the sufficient statistics for $f(x; \bar{X}, s^2) = t_{N-1}$ irrespective of possible values taken on by μ and σ^2. So it would appear, from the point of view of this loss function that preserving the known normality is not critical.

Of course this is all from a frequentist vantage and we may be interested in this problem from a Bayesian viewpoint. For example if we assume a priori

$$(1.5) \qquad g(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

then the predictive distribution of X_{N+1},

$$(1.6) \qquad f(x_{N+1}|x^{(N)}) = \int\int f(x_{N+1}|\mu, \sigma^2)\, p(\mu, \sigma^2|x^{(N)})\, d\mu d\sigma^2$$

is identical to the t_{N-1} confidence distribution of (1.2), where $p(\mu, \sigma^2|x^{(N)})$ represents the posterior density of μ and σ^2.

Presumably if one wants to predict X_{N+1} given the prior assumptions then the t_{N-1} distribution is appropriate. On the other hand, is this a good Bayesian estimator of $N(\mu, \sigma^2)$? From one point of view -- minimizing squared error, the predictive distribution is the pointwise expectation of the sampling distribution. This is clear by taking the integral w.r.t. x_{N+1} of both sides of (1.6) from $-\infty$ to y and interchanging integrals on the r.h.s. Now suppose we do not confine our interest to a single value but focus on several future observations or that fraction of the future set $X_{N+1}, X_{N+2}, \ldots$ that will lie in a certain region. For example if we let $Y_i = X_{N+i} - \bar{X}$, $i=1,\ldots, M$ and $Y' = (X_{N+1}-\bar{X}\ldots, X_{N+M} - \bar{X})$ then Y is $N(0, \Lambda)$ where

$$(1.7) \qquad \lambda_{ij} = \text{Cov } Y_i Y_j = \begin{cases} \sigma^2(1 + N^{-1}) & \text{if } i = j \\ \sigma^2 N^{-1} & \text{if } i \neq j \end{cases}.$$

Now $\frac{(N-1)s^2}{\sigma^2}$ is χ^2_{N-1} so the joint distribution of $Z = \frac{Y}{s\sqrt{1 + N^{-1}}}$ the vector of pivotals, can be shown to be an exchangeable multivariate student distribution independent of the parameters μ and σ for all M. Hence a joint tolerance region for the set $X_{(M)}$ may be computed within the frequentist framework. Let

$$(1.8) \qquad W_i = 1 \quad \text{if} \quad Z_i \in I'$$

$$W_i = 0 \quad \text{if} \quad Z_i \notin I',$$

then $\bar{W} = M^{-1} \sum_{i=1}^{M} W_i$ is the fraction of Z_i's that lie in I'. The chance of the event can be computed in principle from the exchangeable distribution

of Z. If we now let $I = s\sqrt{1+N^{-1}}\, I' + \bar{X}$, than an associated confidence coefficient for the fraction of X's that lie in I may be obtained. We further note that the derived sequence $W_1,\ldots,W_M$ is also exchangeable for all M. Hence it is tempting to see what light, if any, De Finetti's theorem can shed here. This theorem yields the representation

$$(1.9) \qquad \Pr[\bar{W} = \frac{r}{M}] = \binom{M}{r} \int_0^1 \beta^r (1-\beta)^{M-r} f(\beta)\, d\beta$$

for some random variable $\beta = \lim_{M\to\infty} \frac{r}{M}$ with density $f(\beta)$ concentrated on the unit interval. This indicates that $W_1,W_2,\ldots$ can be considered as an i.i.d. Bernoulli sequence conditional on β or

$$(1.10) \qquad \Pr(\bar{W} = \frac{r}{M}|\beta) = \binom{M}{r} \beta^r (1-\beta)^{M-r} .$$

However within the strict frequentist framework under discussion the interpretation of β remains obscure because β is not merely a function of μ and σ^2 but must also depend on $\bar{x}$ and s^2. Clearly the sequence $W_1,W_2,\ldots$ is symmetrically dependent irrespective of μ and σ^2 but is i.i.d. only if $\bar{x}$ and s^2 are fixed in addition. Conditioning on $\bar{x}$ and s^2 however obliterates the pivotal quality of W and Z and renders unavailable the frequentist analysis.

The difficulty is, of course, that the pivotal requires $\bar{X},s,X_{N+1},\ldots,X_{N+M}$ to be random and inference is for repetitions of both past and future while what is desired is inference about the future given the past.

At this point one may further object on the grounds that our presumed sampling of the future is a parametrically conditional i.i.d. situation, thus requiring the fraction of future X's lying in I to converge to $\Pr[X \in I|\mu,\sigma^2]$, the chance that the next observation lies in I.

Again this does not obtain for the sequence of Z's or W's. However since μ and σ^2 are unknown, inference about this chance must in some manner reflect the resulting uncertainty about our knowledge of the sampling distribution function.

These issues, more naturally examined and clarified within a Bayesian context, will be addressed again in section 3.

2. Estimation of the Distribution Function

In the usual Bayesian decision approach we have X the observation space, Θ the parameter space and a the action space with a random variable X having density $f_X(x|\theta)$, $\theta \in \Theta$ and a stipulated loss function $L(a(x), \theta)$ where $a \in a$. Having observed $X = x$ and assumed a prior density $g(\theta)$ for θ, we compute the posterior density of θ,

$$(2.1) \qquad p(\theta|x) \propto f(x|\theta)g(\theta)$$

and the expected loss for each action a

$$(2.2) \qquad L(a) = \int L(a,\theta)p(\theta|x)d\theta.$$

The choice of the appropriate action, say a^*, is obtained from

$$(2.3) \qquad L(a^*) = \min_a L(a) .$$

We apply this now to the estimation of a distribution function. Suppose $X^{(N)} = (X_1, \ldots, X_N)$ are i.i.d. random variables with common distribtuion function $F(x|\theta)$ where θ may be a set of parameters. Let $\gamma = \gamma(y,\theta) = F(y|\theta)$ and let $g(\theta)$ represent a

a prior density for θ. In principle one computes the distribution of γ from the posterior density of θ given $x^{(N)}$, i.e. from

$$(2.4) \qquad p(\theta|x^{(N)}) \propto g(\theta) \prod_{i=1} f(x_i|\theta),$$

obtain $p(\gamma|x^{(N)})$, say for every y. Hence one can compute $1 - \alpha$ probability limits or a point estimator of $\gamma = F(y|\theta)$ for each value of y.

This then would be a Bayesian approach to the estimation of the assumed underlying sampling distribution of the observables. Although for many cases such a procedure could be quite complicated, we shall illustrate a case where it is fairly simple to achieve such an estimation program. Let

$$(2.5) \qquad \gamma = F(y|\theta) = 1-e^{-y\theta}$$

the simple exponential distribution function. Further let $X_1,\ldots, X_d$ be fully observed values while $X_j, j=d+1,\ldots, N$ be censored at values x_j respectively. Suppose we assume a non-informative prior $g(\theta) \propto \theta^{-1}$. There the likelihood is given as

$$(2.6) \qquad L(\theta) = \theta^d \, e^{-\theta N\bar{x}}$$

where $N\bar{x} = \sum_{1}^{N} X_i$ and the posterior density of θ is

$$(2.7) \qquad p(\theta|\bar{x}) = (N\bar{x})^d \, \theta^{d-1} \, e^{-\theta N\bar{x}}/\Gamma(d),$$

or $2\theta N\bar{x}$ is a χ^2 variate with $2d$ degrees of freedom. Now let $\gamma = 1 - e^{-y\theta}$ or $\theta = -y^{-1} \log (1 - \gamma)$ so that the posterior density of γ is

$$(2.8) \qquad p(\gamma|\bar{x}) = \left[\log (1-\gamma)^{-1}\right]^{d-1} (1-\gamma)^{y^{-1}N\bar{x}-1} \left(\frac{N\bar{x}}{y}\right)^d /\Gamma(d) \, .$$

However for most applications we can actually use (2.3) instead of (2.4) because for $0 \leq a < b$ we note that

$$(2.9) \qquad \Pr[a \leq \theta \leq b] = \Pr[1-e^{-ya} \leq \gamma \leq 1-e^{-yb}] \, .$$

So probability limits can be made in this case to depend on (2.3) and highest probability density intervals can, in principle, be computed as a single interval. We now turn to point estimators that minimize loss functions pointwise for each y. It is clear that the median $\tilde{\gamma}$ of γ can be obtained from (2.9) by virtue of the fact that the median $\tilde{\theta}$ of θ is given as

$$(2.10) \qquad \tfrac{1}{2} = [\Pr \theta \leq \tilde{\theta}] = \Pr[\gamma \leq 1-e^{-y\tilde{\theta}}] = P[\gamma \leq \tilde{\gamma}]$$

thus $\tilde{\gamma} = 1 - e^{-y\tilde{\theta}}$ or since $\tilde{\theta} = \dfrac{\tilde{\chi}^2}{2N\bar{x}}$ where $\tilde{\chi}^2$ is the median of a $\chi^2(2d)$ variate. Hence the median estimator of γ is

$$(2.11) \qquad \tilde{\gamma} = 1 - \exp -\left(\frac{y\tilde{\chi}^2}{2N\bar{x}}\right).$$

Here the loss function minimizes absolute error in terms of γ

pointwise for y.

Note that $\chi^2(2d)$ may be read off from a χ^2 table or may be conveniently approximated by

$$\tilde{\chi}^2(2d) \doteq 2d\left[\frac{9d-1}{9d}\right]^3 \tag{2.12}$$

using the Wilson-Hilferty approximation.

A conventional frequentist estimator that substitutes the m.l.e. $\hat{\theta} = d/N\bar{x}$ for θ in γ yields

$$\gamma(\hat{\theta},y) = 1 - \exp(-yd/N\bar{x}) \tag{2.13}$$

and is rather close to $\tilde{\gamma}$ because $\tilde{\chi}^2(2d) \doteq 2d$. In point of fact $\tilde{\chi}^2(2d) < 2d$, which implies that $\tilde{\gamma} < \gamma(\tilde{\sigma},y)$ for every y with $\tilde{\gamma} \to \gamma(\hat{\sigma},y)$ as d grows.

A second estimator $\bar{\gamma} = E(\gamma)$, which minimizes squared error in terms of γ pointwise for y, may be computed as

$$\bar{\gamma} = \int \gamma(\theta,y)\, d\, P(\theta|x) = \int_0^y \frac{d(N\bar{x})^d}{(N\bar{x}+x)^{d+1}} dx \tag{2.14}$$

which is obviously the predictive distribution of a future observation drawn from this process. Thus this has a second interpretation.

The modal estimator γ_m, which minimizes in the limit a loss which is proportional to the length of the interval if correct and a constant less if incorrect, is obtained from maximizing the density $p(\gamma|\bar{x})$. This results in

$$(2.15)\quad \gamma_m = 0 \qquad \text{for } y < N\bar{x},\ d = 1$$

$$= \tfrac{1}{2} \text{ (by definition)} \qquad \text{for } y = N\bar{x},\ d = 1$$

$$= 1 - \exp\left[-\frac{(d-1)y}{N\bar{x} - y}\right] \qquad \text{for } y \leq N\bar{x},\ d > 1$$

$$= 1 \qquad \text{for } y > N\bar{x},\ d \geq 1$$

Although the estimator γ_m assumes an extreme value when $y \geq N\bar{x}$, a simple analysis of the situation indicates that this fact is not surprising. The shortest interval on γ for any $y \geq N\bar{x}$ has a terminus at $\gamma = 1$ i.e.

$$(2.16)\quad \Pr[1-e^{-ya} \leq \gamma \leq 1] = 1 - \alpha \quad \text{for } y \geq N\bar{x}$$

and yields an interval of size e^{-ya}, the shortest possible one in this situation for the associated α. On the other hand in the case $y < N\bar{x}$ but $d = 1$, the shortest interval for the associated α is rendered by

$$(2.17)\quad \Pr[0 \leq \gamma \leq 1-e^{-yb}] = 1 - \alpha$$

The case $d = 1$ is the most interesting as here γ_m is 0 or 1 depending on y being less than or exceeding $N\bar{x}$, and defined as $\frac{1}{2}$ for $y = N\bar{x}$.

$$(2.18)\quad p(\theta|\bar{x}) = N\bar{x}\, e^{-\theta N\bar{x}}$$

$$(2.19)\quad 1-e^{-bN\bar{x}} = P(\theta \leq b) = P(\gamma \leq 1-e^{-yb}) \quad \text{for } y < N\bar{x}$$

For $y > N\bar{x}$, and the same probability

$$(2.20)\quad 1-e^{-bN\bar{x}} = P(\theta \geq a) = P(\gamma \geq 1-e^{-ya}) = \Pr[\gamma \geq 1 - (1-e^{-bN\bar{x}})^{y/N\bar{x}}]$$

since

$$a = \frac{1}{N\bar{x}} \log(1-e^{-bN\bar{x}})^{-1}$$

and

$$1-e^{(-y/N\bar{x})\log(1-e^{-bN\bar{x}})} = 1 - (1-e^{-bN\bar{x}})^{y/N\bar{x}}$$

At $y = N\bar{x}$ all $1 - \alpha$ intervals have the same length due to γ being uniformly distributed.

An illustration of intervals of this type is given in Figure 3.

3. The Predictive Approach

This approach essentially has a different goal, choosing an action which depends on the value of a future observable rather than on the value of the parameter. Here we postulate the sampling density of future $X_{(M)} = (X_{N+1}, \ldots, X_{N+M})$ random variables conditional on θ and on a current set of random variables $X^{(N)}$ to be

$$(3.1) \qquad f(x_{(M)} | x^{(N)}, \theta).$$

Instead of a parametric loss function we posit a predictive loss function c.f. Aitchison and Dunsmore (1975),

$$(3.2) \qquad L_p(a, x_{(M)}).$$

We then obtain the predictive density of the future $X_{(M)}$ given the current data $X^{(N)} = x^{(N)}$ as

$$(3.3) \qquad f(x_{(M)} | x^{(N)}) = \int f(x_{(M)} | x^{(N)}, \theta)\, p(\theta | x^{(N)}) d\theta$$

and average predictive loss

$$L_p(a) = \int L_p(a, x_{(M)})\ f(x_{(M)}|x^{(N)})\ dx_{(M)}.$$

Now we choose a^* such that

$$(3.4) \qquad L_p(a^*) = \min_a L_p(a).$$

If we again consider the i.i.d. case then it is easily shown under fairly general conditions that there exists a

$$(3.5) \qquad L(a,\theta) = \int L_p(a, x_{(M)})\ f(x_{(M)})|x^{(N)}, \theta)\ d\ x_{(M)}$$

for every $L_p(a, x_{(M)})$ although the converse need not be true.

In the case of i.i.d. random variables X_i, $i = 1, \ldots, N+M$ having common distribution function $F_X(x|\theta)$, the predictive density of

$$(3.6) \qquad f(x_{N+1}, \ldots, X_{N+M}|x^{(N)}) = \int \left[\prod_{i=1}^{M} f(x_{N+i}|\theta) \right] p(\theta|x^{(N)})\ d\theta$$

represents a set of M exchangeable random variables (rarely independent). This results in an apparent distinction between the probability that single future observation lies in some set I and the fraction of all such future observations that lie in I. Before we delve into that question we first indicate under what conditions it is appropriate to consider only the marginal distribution of a single future observation. Suppose that our loss function is of the additive form with equal loss for each component,

$$(3.7) \qquad L_p^{(M)}(a, x_{(M)}) = \sum_{i=1}^{M} L_p(a, x_{N+i}),$$

then the average loss,

$$(3.8) \qquad L_p^{(M)}(a) = \sum_{i=1}^{M} \int\cdots\int L_p(a, x_{N+i})\, f(x_{N+1},\ldots, x_{N+M}|x^{(N)})\, dx_{(N+1},\ldots, dx_{N+M}$$

$$= M \int L_p(a,x)\, f(x|x^{(N)})\, dx = ML_p(a)$$

where $f(x|x^{(N)})$ represents the common marginal density of the exchangeable set of future random variables $X_{N+1},\ldots, X_{N+M}$. Hence the average loss depends only on the marginal distribution.

Given suitable conditions which permit interchanging infinite iterated integrals and an infinite sum then this remains true as $M \to \infty$.

Now as an example consider a loss function of the type

$$(3.9) \qquad L(I,x) = \begin{cases} \delta\, m(I) & \text{if } x \notin I \\ \delta\, m(I) - K & \text{if } x \in I \end{cases}$$

for $K > 0$ where I is a countable set of non-overlapping intervals, and $m(I)$ is total length of I. Then

$$(3.10) \qquad L(I^*) = \min_{I} L(I)$$

yields $I^* = \{x; f(x|x^{(N)}) > \frac{\delta}{K}\}$.

Hence it is sensible, first of all, to consider the same set I, independently of the number of future observations under consideration. Let us now consider the derived random variable

$$(3.11) \qquad Y_i = \begin{cases} 1 & \text{if } X_{N+i} \in I \\ 0 & \text{otherwise} \end{cases}$$

for any particular measurable set I.

Then $Y_1, \ldots, Y_M$ are exchangeable so that

$$(3.12) \qquad E(\bar{Y}) = Pr(X_{N+i} \in I) = \int_I f(x_{N+i} | x^{(N)})\, dx = Pr(Y_i = 1) = q$$

where $\bar{Y} = M^{-1} \sum_{i=1}^{M} Y_i$ and

$$(3.13) \qquad Var(\bar{Y}) = q(1-q)\left(\frac{1}{M} + \frac{M-1}{M}\rho\right) = q(1-q)\left(\frac{1-\rho}{M} + \rho\right)$$

where for all $i \neq j$,

$$(3.14) \qquad \rho = \frac{Pr\,[Y_i=1,\ Y_j=1] - q^2}{q(1-q)}$$

Note for $\rho \geq 0$, which must clearly hold, otherwise $Var(\bar{Y})$ will not be non-negative, that

$$(3.15) \qquad \lim_{M \to \infty} (Var\ (\bar{Y}) = \rho q(1-q) > 0.$$

Further ρ is a function only of N and not M, that tends to zero for increasing N under fairly general conditions, i.e., the conditions under which $p(\theta | x^{(N)})$ tends to concentrate all of its mass at a single point. But for any fixed N, the uncertainty in the fraction, $\bar{Y}$, of all future observations that lie in I, does not in general go to zero for finite N. It is true that its expectation is exactly the chance that $X_{N+i} \in I$ as given by the marginal predictive distribution of X_{N+i}.

Superficially, at least, one would think that $\bar{Y}$ should

converge in probability to its mean value $q = Pr(X_{N+i} \in I)$ as M increases as it obviously would if the sequence $X_{N+1}, X_{N+2}, \ldots$ were additionally independent instead of just exchangeable. However in point of fact in our original model they are considered to be independent conditional on θ. These points are not really discordant in the sense that clearly conditional on θ, $\bar{Y}$ converges to $Pr(X_{N+i} \in I|\theta)$ and the fact that it doesn't unconditionally is a reflection of our uncertainty about the actual sampling distribution $F(x|\theta)$. It would be a paradox of the Bayesian procedure if unconditionally $\bar{Y}$ did converge since the interpretation would be that $F(x_{N+i}|x^{(N)})$ was actually the sampling distribution instead of the predictive distribution. At any rate, as the reader may already have recognized, this is merely a consequence again of De Finetti's famous theorem where he shows, in terms of our situation that $\bar{Y}$ converges to $\gamma = P(X \in I|\theta)$ with density $p(\gamma|x^{(N)})$, using γ in a wider sense than previously.

In general then for finite M and $r = 0, 1, \ldots, M$

$$(3.16) \qquad Pr\left[\bar{Y} = \frac{r}{M}\right] = \int \binom{M}{r} \gamma^r (1-\gamma)^{M-r} \, p(\gamma|x^{(N)}) \, d\gamma.$$

In particular for the simple exponential example we have for $\gamma = 1-e^{-\theta y}$,

$$Pr\left[\bar{Y} = \frac{r}{M}\right] = \binom{M}{r} (N\bar{x})^d \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} \, [N\bar{x}+(M-j)y]^{-d}$$

and of course $\lim_{M\to\infty} \bar{Y} = \gamma$ which has density $p(\gamma|x^{(N)})$ as derivable from (2.8).

4. Illustrations

We illustrate the use of the estimators with some Russian data obtained from Gnedenko, Belyayev and Solovyev (1969, p. 176), and then some artificially generated data. For the Russian data a sample of 100 items are tested and time to failure recorded for each up until 500 standard time units have elapsed. During this period 89 items have survived and the recorded failure times for the other 11 are: 31, 49, 90, 135, 161, 249, 323, 353, 383, 436, 477. The total time on test is $47,187 = N\bar{x}$.

The three estimates of the assumed exponential survival distribution function are plotted in Figure 1 along with 90% posterior limits - 5% at each tail.

[Place Figure 1 here]

Two artificially generated data sets from an exponential distribution with parameter $\theta = 1$ are also presented. In the first set of 8 observations 4 were recorded as .371, .525, .033, .027, and 4 exceeded the assigned censored value .75. The three estimates of the distribution function and 90% posterior limits - 5% at each tail are plotted in Figure 2.

[Place Figure 2 here]

The second data set is to illustrate the case of $d = 1$, where the modal estimator is a single step function. Four observations were generated from the previous distribution with censoring now at .25. This yielded the single observed value of .059 with the other three exceeding .25. The same plot as before is presented in Figure 3 except that now the 90% posterior limit is one-sided because here the highest posterior density limits are easily obtained because of the monotonic nature of the density of γ for each y when $d = 1$. All values of γ between 0 and the

upper curve are included in the 90% interval up to the point $y = N\bar{x} = .809$. Beyond this point the 90% interval includes all values of γ above the lower curve. At $y = .809$, γ is uniformly distributed and the 90% limits are arbitrarily set at (.05, .95) although any interval of length .9 would suffice.

[Place Figure 3 here]

5. Remarks

When and how does one use these results for prediction? Clearly, for the Bayesian who wants only to predict the next observation -- the predictive distribution is appropriate. However, if one wants to estimate (predict) the fraction of all future observations lying in some prescribed set then the estimating apparatus is sensible. But note that one reasonable point-wise Bayesian estimator is the probability derived from the predictive distribution alluded to above. For the case involving a fraction of a finite number of future values, the appropriate predictive distribution is given in (3.16). This is easily calculable when M is not very large and should also be approximately manageable for large M due to its convergence. Further the expected value of the fraction is the same as the probability of the next observation lying in that set so that in any event, the preeminence of the predictive distribution is clear.

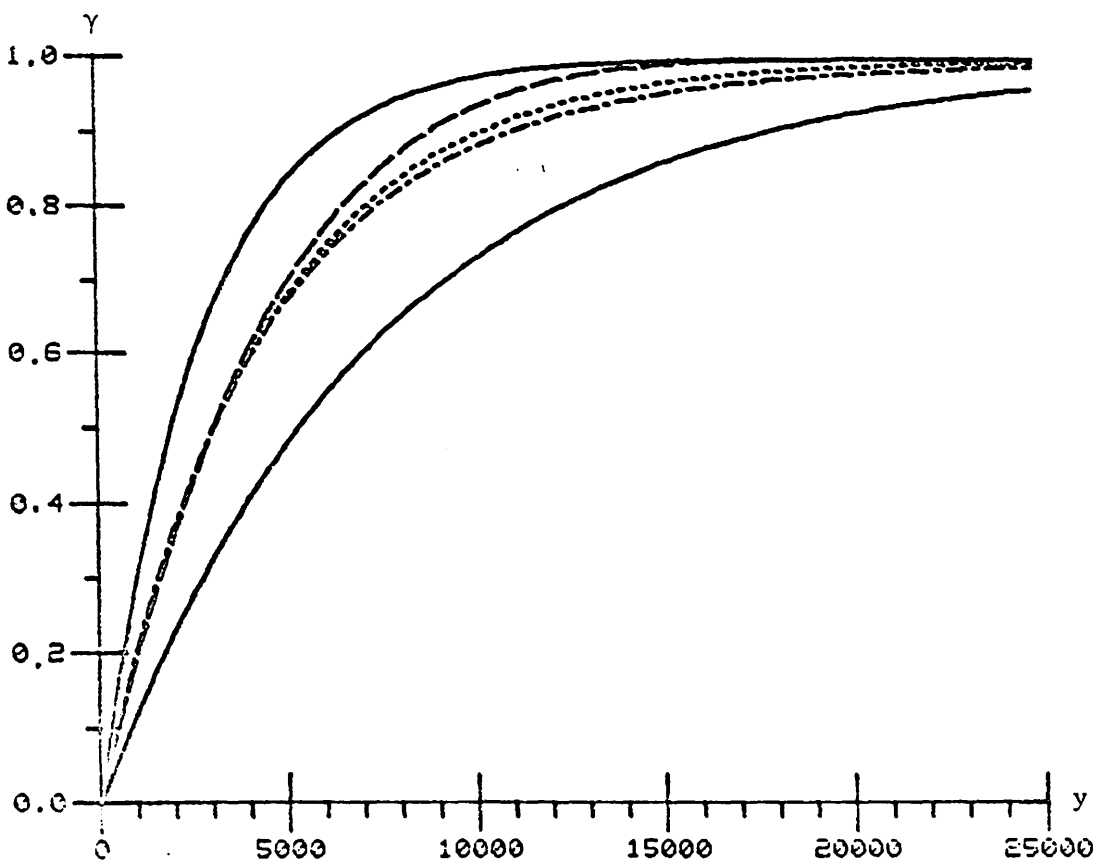
The example chosen here -- exponential survival with censoring -- has the dual virtue of clarity of interpretation and mathematical simplicity. In contrast, a frequentist analysis must attend to the stopping rule, wherein either one or both of the number and time of censoring are random variables, with the consequence that each of the several possible sampling situations entails a distinct analysis. The details of the various methods in estimating θ are presented in Gnedenko et al (1969). Even greater

complexity ensues in the construction of tolerance regions for one or more future observables.

Approaches, not altogether different from that taken here, were presented in a series of papers in discriminatory analysis, Desu and Geisser (1973), Enis and Geisser (1971, 1972), Geisser (1967, 1970, 1977), and in estimating the chance that one random variable exceeds another, Enis and Geisser (1970). There they were often called Semi-Bayes procedures.

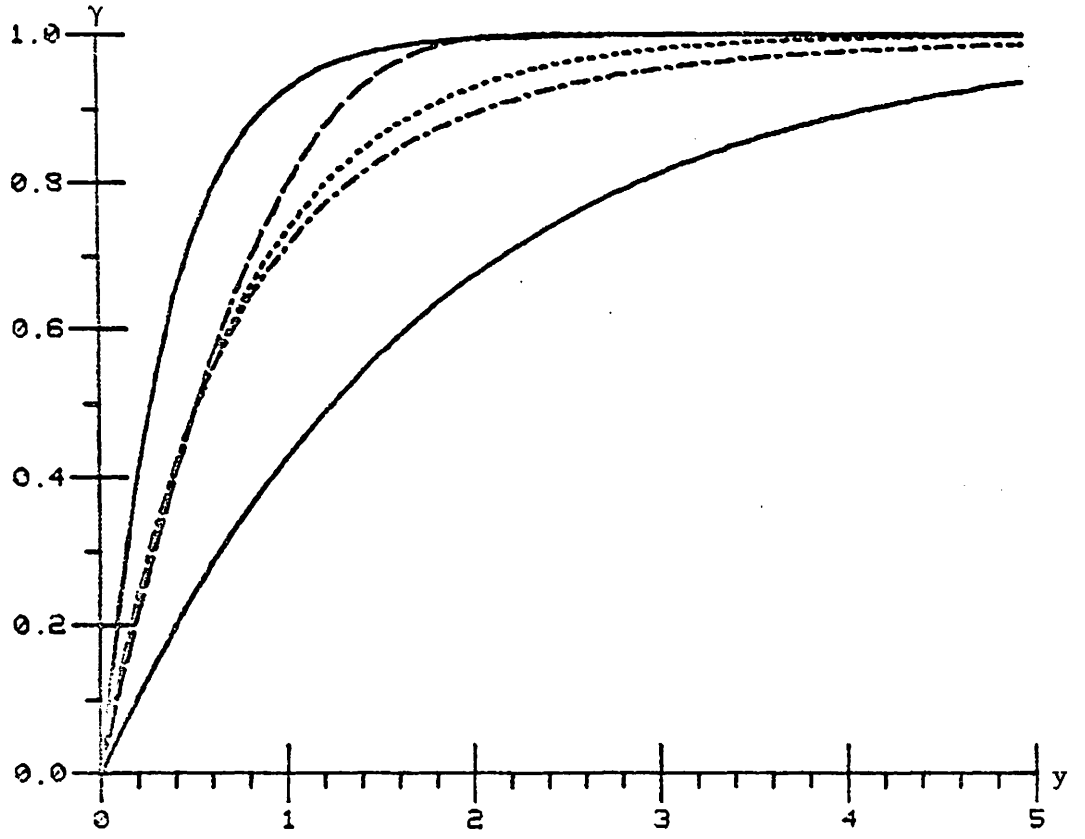
I am indebted to Dennis Jennings for the computations and graphs of Figures 1-3.

Figure 1. Graph of estimates and probability limits on $\gamma = F(y|\theta)$ from Russian data.



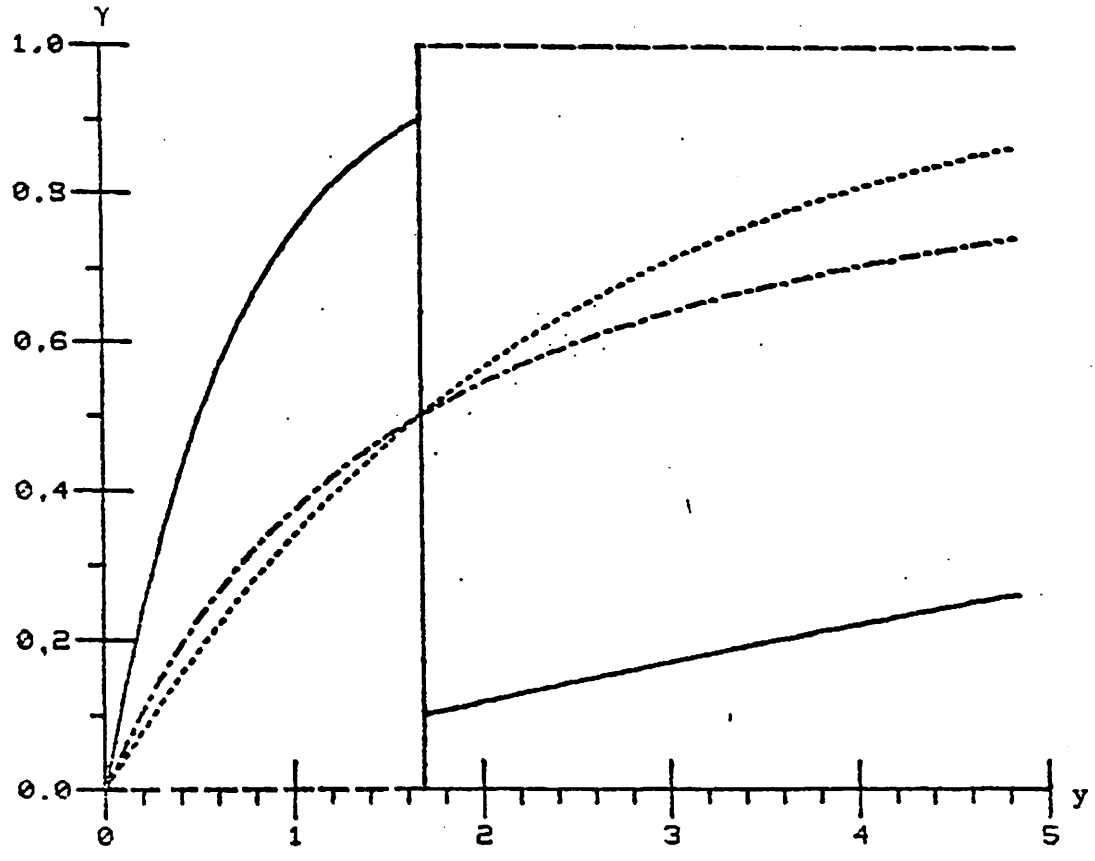
Legend:
- modal estimate γ_m
- median estimate $\tilde{\gamma}$
- mean estimate $\bar{\gamma}$
- upper and lower 5% limits on γ

Figure 2. Graph of estimates and probability limits on $\gamma = F(y|\theta)$ from 8 generated observations



Legend:
- modal estimate γ_m
- median estimate $\tilde{\gamma}$
- mean estimate $\bar{\gamma}$
- upper and lower 5% limits on γ

Figure 3. Graph of estimates and probability limits on $\gamma = F(y|\theta)$ from 4 generated observations



Legend:
- modal estimate γ_m
- median estimate $\tilde{\gamma}$
- mean estimate $\bar{\gamma}$
- shortest 90% limits

References

Aitchison, J. and Dunsmore , I. R. (1975). Statistical Prediction Analysis Cambridge University Press, London.

Desu, M. M. and Geisser, S. (1973), Methods and applications of equal-mean discrimination, Discriminant Analysis and Applications, edited by T. Cacoullos, Academic Press, New York, pp. 139-161.

Enis, P. and Geisser, S. (1970), Sample discriminants which minimize posterior squared error loss, South African Statist. J., 4, pp.85-93.

Enis, P. and Geisser, S. (1971), The estimation of the probability that $Y < X$, Journal of the Amer. Stat. Assoc., 66, 333, pp. 162-168.

Enis, P. and Geisser, S. (1974), Optimal predictive linear discriminants, Ann. Statist., 2, 2, pp. 403-410.

Geisser, S. (1967), Estimation associated with linear discriminants, Ann. Math. Statist., 38, pp. 807-817.

Geisser, S. (1970), Discriminatory practices, Bayesian Statistics, edited by D. Meyer and R. C. Collier, Peacock, Illinois, pp. 57-70.

Geisser, S. (1977), Discrimination, allocatory and separatory, linear aspects, Classification and Clustering, edited by J. Van Ryzin, Academic Press, New York, pp. 301-330.

Gnedenko, B. B., Belyayev, Yu. K. and Solovyev, A. D. (1969). Mathematical Methods of Reliability Theory. Academic Press, New York and London.

Kullback, S. and Leibler, R. A. (1951), On information and sufficiency, Annals of Math. Stat., 22, pp. 79-86.

Murray, Gordon D. (1977), A note on the estimation of probability density functions, Biometrika, 64, pp. 150-152.